DATA SCIENCE FOR BUSINESS

Managing data project with quinten

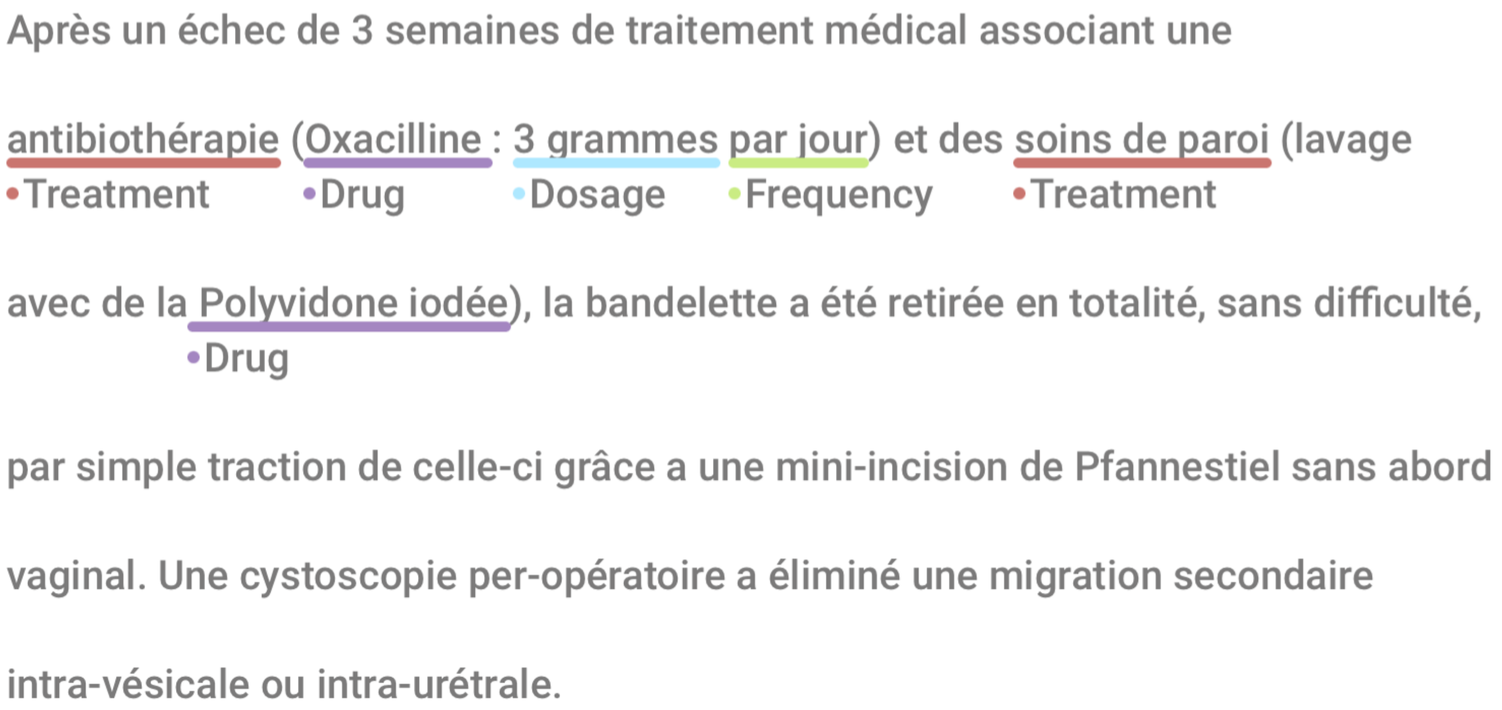
Named-entity recognition & data augmentation

Group 1
Pierre de Boisredon
Sylvain Delgendre
Clément Girault
Thomas Kessous
Adrien Salvador
François Schmerber

ANNOTATION

with Doccano

Labels: Treatment, Drug, Dosage, Frequency, Form, Duration



Après un échec de 3 semaines de traitement médical associant une antibiothérapie (Oxacilline : 3 grammes par jour) et des soins de paroi (lavage avec de la Polyvidone iodée), la bandelette a été retirée en totalité, sans difficulté, par simple traction de celle-ci grâce a une mini-incision de Pfannestiel sans abord vaginal. Une cystoscopie per-opératoire a éliminé une migration secondaire intra-vésicale ou intra-urétrale.

DATA AUGMENTATION

Back Translation

Mr. R, **âgé de** 53 ans sans **antécédents** pathologiques particuliers a été admis aux urgences pour **rétention aiguè d'urine** avec hématurie macroscopique

↓ Back Translation without labeled words

Mr.R, 53 ans sans **histoire** pathologique particulière a été admis **en cas** d'urgence pour **rétention d'urine aiguë** avec **une** hématurie macroscopique

Le patient a été **mis** sous héparine **de bas** poids moléculaire et antibiothérapie.

↓ Back Translation, labelled words remain unchanged

Le patient a été **placé** sous héparine **faible** poids moléculaire et antibiothérapie.

DATA AUGMENTATION

Back Translation

Strategy: Translating in english then in French the sequence of words which do not contain labelled words

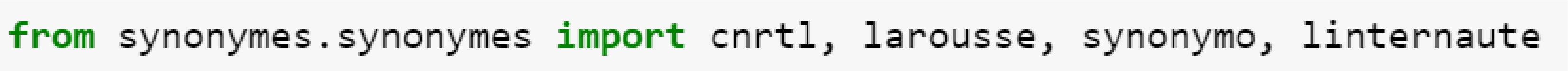
+

- Words are translated in their context
- English and French are quite similar so the quality of translation is satisfactory most of the time

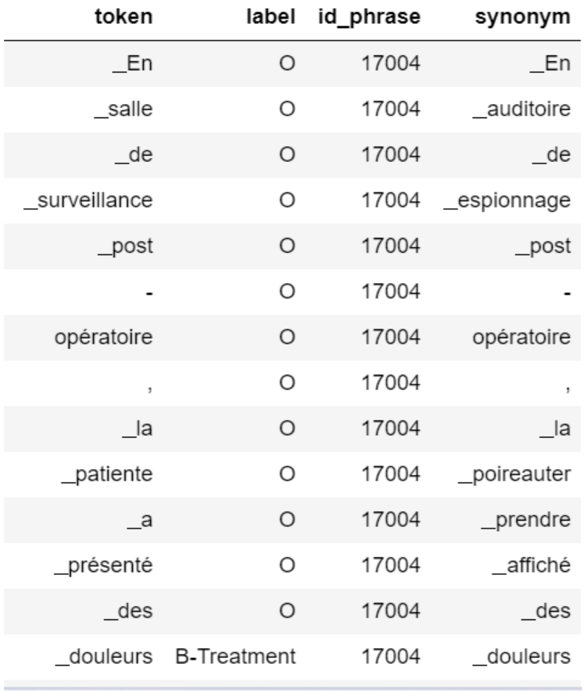
−

- Medical words can be mistranslated:
  - transurétrale → transureuse
- Punctuation is sometimes added:
  - Monsieurb**,** → Monsieurb**.,**

Data Augmentation: synonyms replacement



```
from synonymes.synonymes import cnrtl, larousse, synonymo, linternaute
```



token	label	id_phrase	synonym
_En	O	17004	_En
_salle	O	17004	_auditoire
_de	O	17004	_de
_surveillance	O	17004	_espionnage
_post	O	17004	_post
-	O	17004	-
opératoire	O	17004	opératoire
,	O	17004	,
_la	O	17004	_la
_patiente	O	17004	_poireauter
_a	O	17004	_prendre
_présenté	O	17004	_affiché
_des	O	17004	_des
_douleurs	B-Treatment	17004	_douleurs

The Choice of the library is important…

or this can happen:

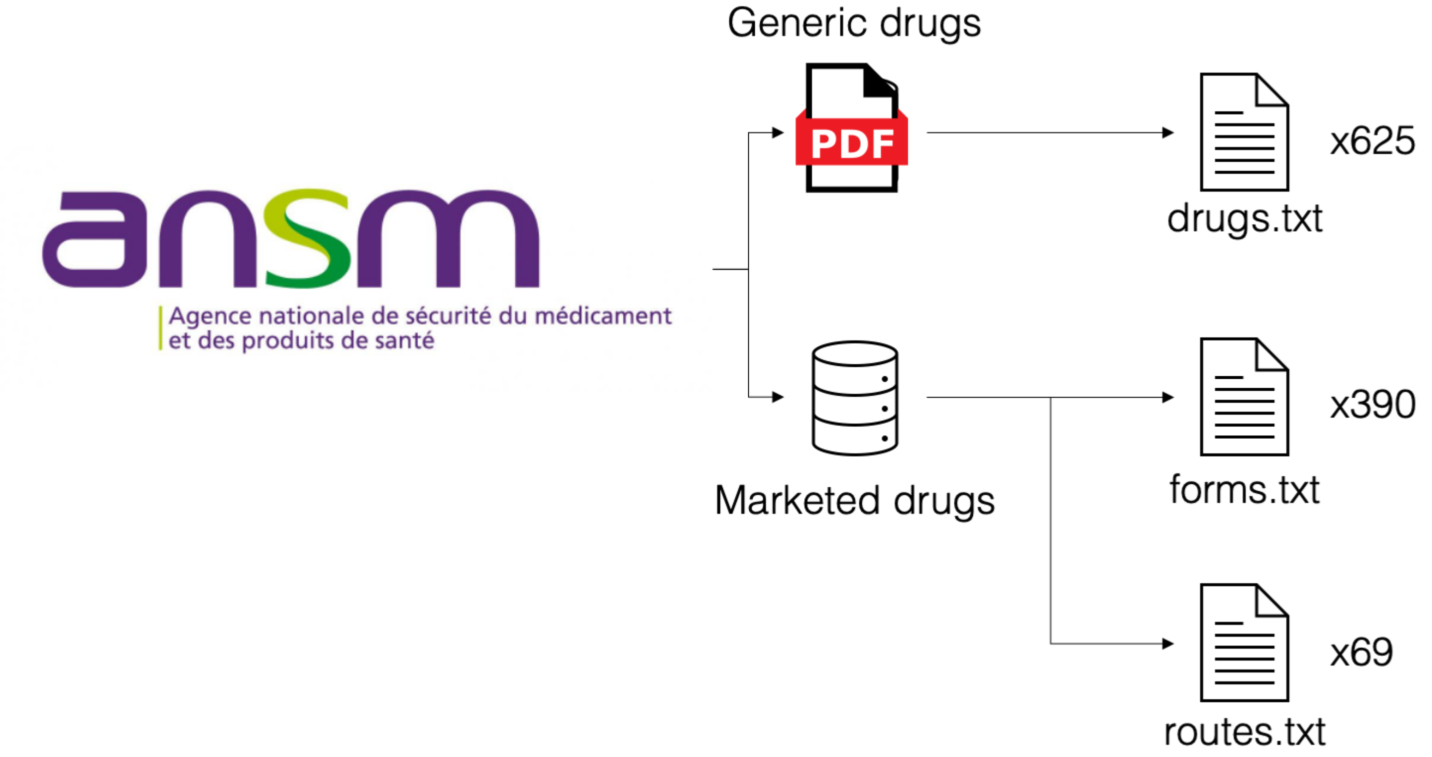
“patiente”

~“patienter”

“poireauter”

DATA AUGMENTATION

External data



TOKENIZATION

with CamemBERT

```
'Pyrazinamide 1500 mg
```

Drug Dosage

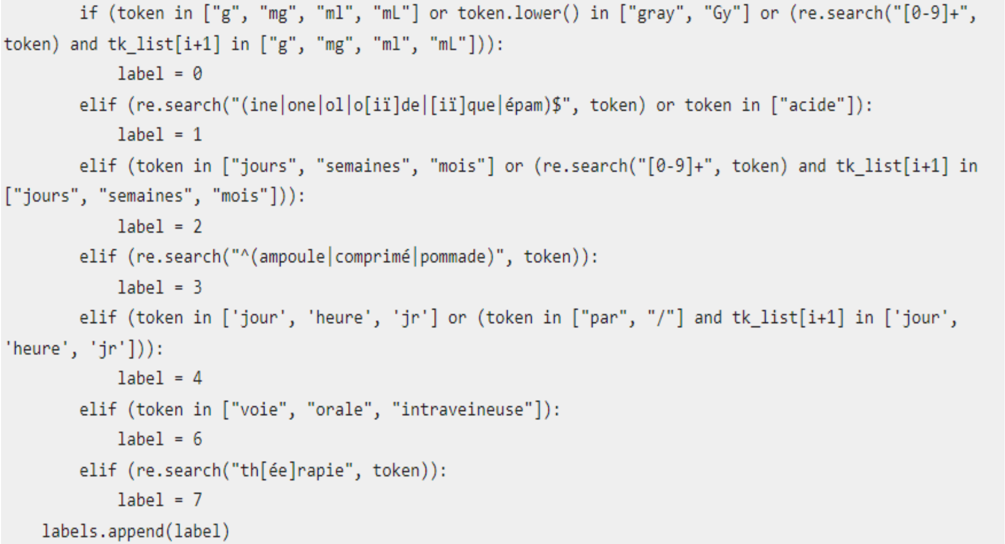
```
le matin Arrêt'
```

CamemBERT Tokenizer

```
['▁Pyr',        ['B-Drug',
'a',            'I-Drug',
'zin',          'I-Drug',
'ami',          'I-Drug',
'de',           'B-Dosage',
'▁1500',        'I-Dosage',
'▁mg',          'O',
'▁le',          'O',
'▁matin',       'O',
'▁Arrêt']       'O']
```

MODEL

Regex - Baseline



```
        if (token in ["g", "mg", "ml", "mL"] or token.lower() in ["gray", "Gy"] or (re.search("[0-9]+",
token) and tk_list[i+1] in ["g", "mg", "ml", "mL"])):
            label = 0
        elif (re.search("(ine|one|ol|o[iï]de|[iï]que|épam)$", token) or token in ["acide"]):
            label = 1
        elif (token in ["jours", "semaines", "mois"] or (re.search("[0-9]+", token) and tk_list[i+1] in
["jours", "semaines", "mois"])):
            label = 2
        elif (re.search("^(ampoule|comprimé|pommade)", token)):
            label = 3
        elif (token in ['jour', 'heure', 'jr'] or (token in ["par", "/"] and tk_list[i+1] in ['jour',
'heure', 'jr'])):
            label = 4
        elif (token in ["voie", "orale", "intraveineuse"]):
            label = 6
        elif (re.search("th[ée]rapie", token)):
            label = 7
    labels.append(label)
```

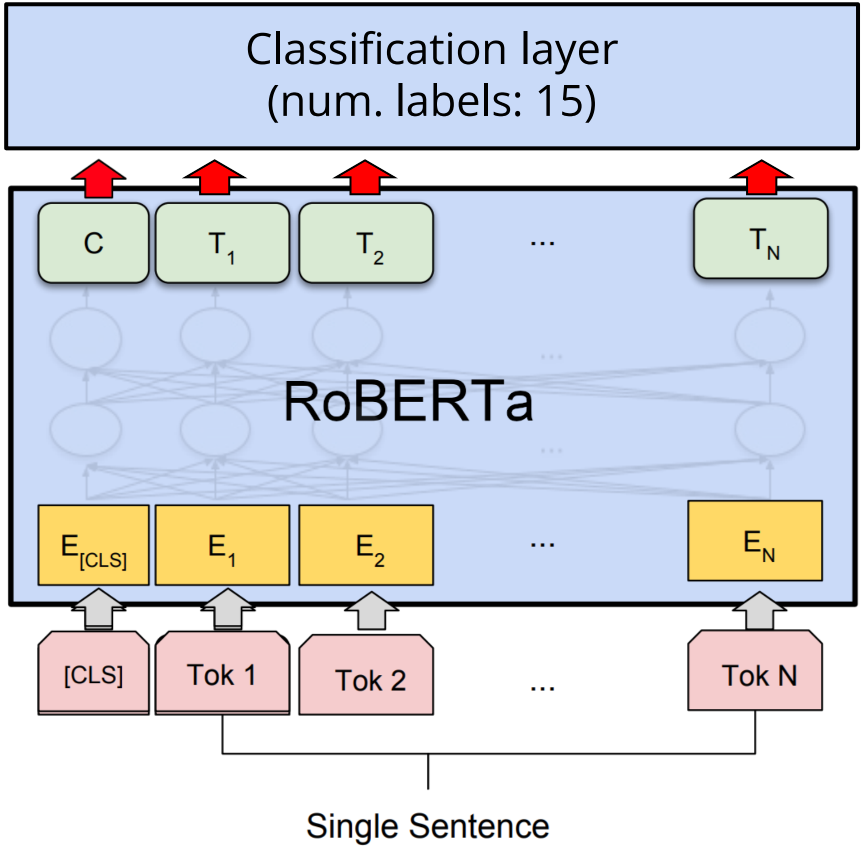
Comments:

- Predict label if specific pattern is detected in the text
- Do not use information about the full sentence
- No learning

➢ Used as a baseline model to benchmark our final prediction

MODEL

Final model



Hyperparameters:

Batch size: 32
Epochs: 16
Learning rate: 5e-5

Model

Results

Regex

Score: 0.52
Training time: 0s
Prediction time: 0.2s
no GPU needed

NER - Hugging Face + pyTorch

Score: 0.56
Training time: 15min on GPU
Prediction time: 20s

NER - Hugging Face + pyTorch + regex

Score: 0.632

Conclusion

Result

The model which performs the best is our NER model with transformers, with a final Regex layer

Next steps

- Add other languages / websites to the back translation module
- Add brand names in addition to generic names in our external data
- Try other data augmentation methods